

VictoriaFalls:
Scaling Highly-Threaded Processor Cores

STEPHEN PHILLIPS
Distinguished Engineer, Sun Microsystems
08/21/07

Legal Notice



THESE MATERIALS ARE PROVIDED BY THE COPYRIGHT HOLDERS AND OTHER CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS (INCLUDING ANY OF OWNER'S PARTNERS, VENDORS AND LICENSORS) BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THESE MATERIALS, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Sun, Sun Microsystems, the Sun logo, Solaris, OpenSPARC T1, OpenSPARC T2 and UltraSPARC are trademarks or registered trademarks of Sun Microsystems, Inc. in the U.S. and other countries. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. in the U.S. and other countries. Products bearing SPARC trademarks are based upon architecture developed by Sun Microsystems, Inc. UNIX is a registered trademark in the U.S. and other countries, exclusively licensed through X/Open Company, Ltd. The Adobe logo is a registered trademark of Adobe Systems, Incorporated. Part of the products covered by these materials may be derived from the Berkeley BSD systems licensed by the University of California. Sun Microsystems, Inc. has intellectual property rights relating to technology embodied in the product described in these materials. This distribution may include materials developed by third parties who have intellectual property rights therein. Products covered by and information contained in these materials may be controlled by U.S. Export Control laws and may be subject to the export or import laws in other countries. Nuclear, missile, chemical biological weapons or nuclear maritime end uses or end users, whether direct or indirect, are strictly prohibited. Export or reexport to countries subject to U.S. embargo or to entities identified on U.S. export exclusion lists, including, but not limited to, the denied persons and specially designated nationals lists may be prohibited.

Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

Outline

- Highly-Threaded Processors
- VictoriaFalls Overview
- Scaling Challenges & Implementation
- Performance Scaling
- Summary

Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

Highly-Threaded Processors

- Optimize for Throughput and Application Parallelism
 - > Most Important Commercial Server Applications are Heavily Threaded
 - > Parallelism through Aggregation (Multi-instance, Multi-process)
 - > Virtualized Server Environments (e.g. Logical Domains)
- Attack the Memory Wall
 - > Commercial Workloads Exhibit Poor Memory Locality
 - > For a Single Thread, Memory Latency is the Bottleneck to Improving Performance
 - > Diminishing Returns with Increased Cache Sizes in Terms of Both Performance and Die Area Efficiency
- Trade off Thread Latency for Thread Throughput
 - > For a Single Thread, Only Modest Throughput Speedup is Possible By Reducing Compute Time (Increased Frequency, ILP)

Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

Highly-Threaded Processors

- Architected as Core-centric Designs to Maximiz Thread Count within Die Area Limits
 - > Relatively High Thread Count per Core
 - > Many Cores Imply Small Cores and Associated L1 Caches
 - > Modest Capacity Shared Outermost Cache (L2$)
- Managing High Concurrency at All Levels of the Design is the Major Scaling Challenge
 - > Core and Core-L2$ Interconnect
 - > L2$
 - > Memory & Multi-chip Interconnect

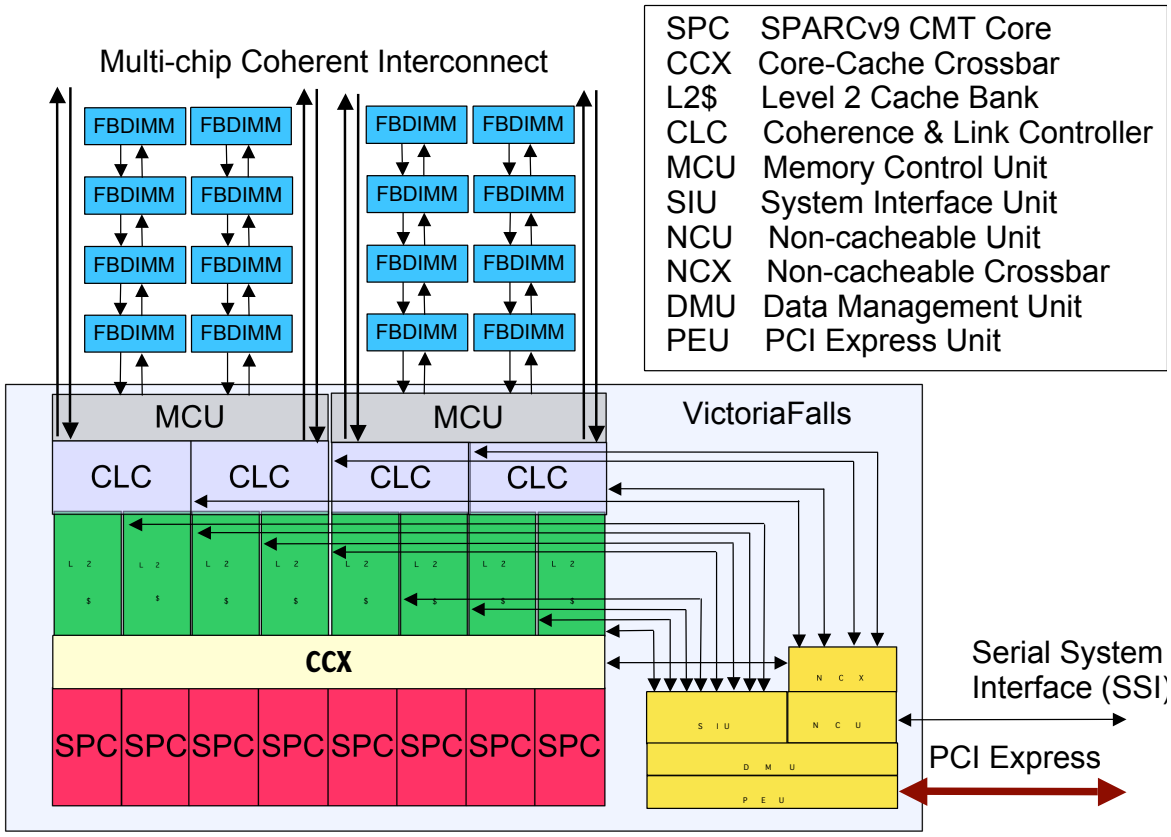
Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

VictoriaFalls Chip Multiprocessor

- 8 Core CMP with 8 Strands per Core @ 1.4Ghz
- Niagara2 SPARCv9 Core
 - > 2 x 8-stage Integer Units (4 Threads per Pipe) – Single Issue per Pipe
 - > 12-stage Pipelined FGU (except divide/sqrt)
 - > Integrated Crypto Accelerator
 - > 16KB 8-way SA L1-I$, 32B Lines, Write-through
 - > 8KB 4-way SA L1-D$, 16B Lines, Write-through
 - > 64 Entry Fully Associative I-TLB
 - > 128 Entry Fully Associative D-TLB
- 4 MB Shared L2$
- 2 Dual-Channel FBDIMM Memory Controllers
- Integrated PCI Express I/O Bridge
- Multi-chip Coherence Links

Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

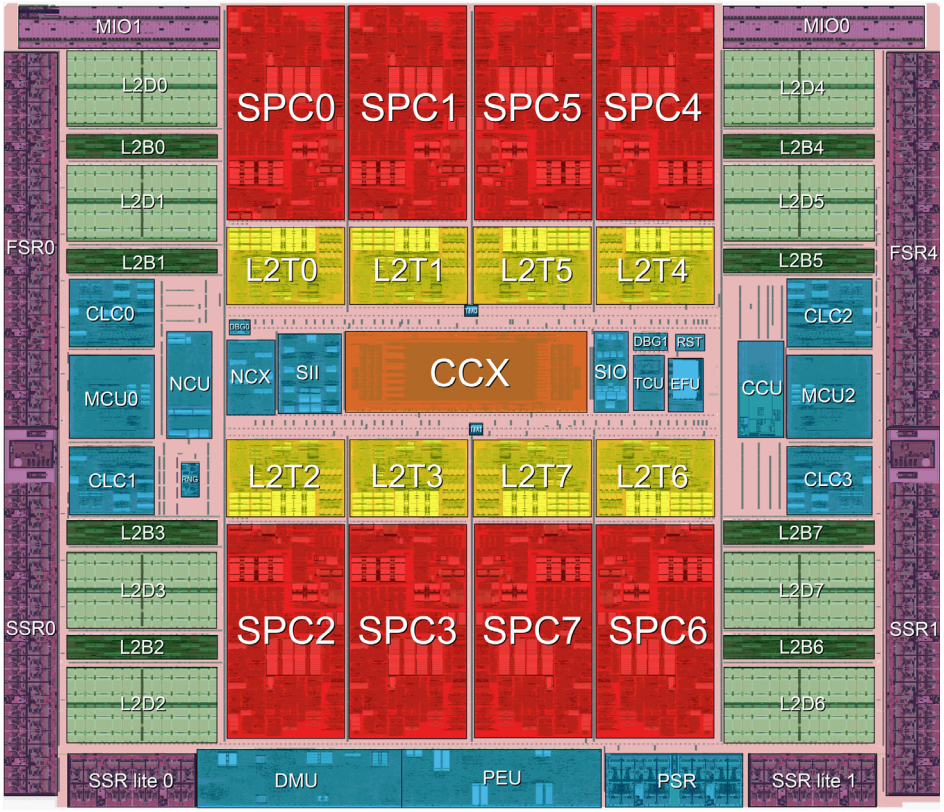
High Level Block Diagram



SPC	SPARCv9 CMT Core
CCX	Core-Cache Crossbar
L2$	Level 2 Cache Bank
CLC	Coherence & Link Controller
MCU	Memory Control Unit
SIU	System Interface Unit
NCU	Non-cacheable Unit
NCX	Non-cacheable Crossbar
DMU	Data Management Unit
PEU	PCI Express Unit

Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

VictoriaFalls Micrograph



65nm CMOS

11 Metal Laye

709 Signal I/

1831 Total I/(

~Niagara2 CM Power

~Niagara2 CM Die Area

Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

1st Level Scaling: Core & L2$ Crossbar

- SPC Memory Model
 - TSO Compliance Maintained by Combination of Core Load/Store Unit and the L2$ Crossbar
 - Exceptions are Instructions That Need not Support TSO
- Non-blocking Core-L2$ Crossbar
 - Establishes Memory Order between Transactions from the Same and Different L2$ Banks
 - Guarantees Delivery of Transactions to L2$ Banks in the Same Order
 - 180GB/sec Hit Bandwidth
- Concurrency
 - 8 Deep Store Buffer per Thread
 - Up to 384 Transactions per Processor (L2$ Limit)

TSO: Total Store Order

Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

2nd Level Scaling: L2$

- 4MB Shared L2$
 - 8 Banks with Independent L2$ Pipelines
 - 64B Cache Lines, Writeback, Modified-LRU Replacement
 - Maintains Directory of L1$ Tags
- Minimizing L2$ Conflicts
 - Limited Thread Speculation
 - High Set Associativity (16-ways per Bank)
 - Set Index Hashing
- Micro-parallelization Challenge
 - Divide Serial Code Segments Among Concurrent Worker Threads
 - Lock-free Synchronization Primitives
 - Reduced Synchronization Overheads Communicating Through Common L2$ Optimized for Core-to-Core Communication
 - No L1-D$ Probes due to Hot Locks, Real or False Data Sharing

Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

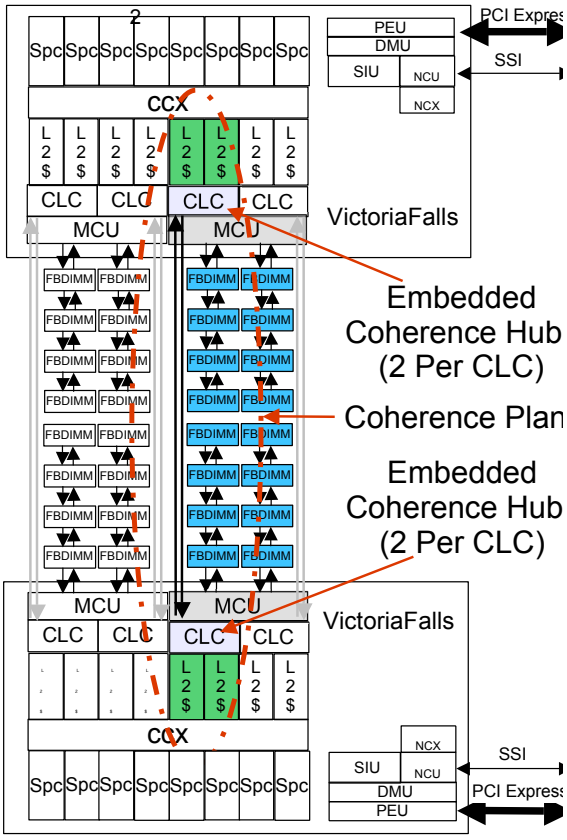
3rd Level Scaling: Memory & Multi-chip

- Memory Bandwidth
 - > Result of Fine-grain Vertical Threading (Niagara2 Core), Core Count and Modest L2$ Size
 - > 50%-80% Core Pipeline Utilization Typical on Commercial Workloads
 - > Commercial Workloads Can Exceed 15GB/sec Average Memory Bandwidth on a Single CMP
 - > Requires Multiple On-die Memory Controllers with Concurrent Read/Write Channel Scheduling
- Snoopy-based Coherent Interconnect
 - > Physical Addresses Partitioned Across 4 Coherence Planes
 - > Multiple Busses Replaced with Point-to-point Links
 - > Direct Chip-to-chip or Hub/Star Physical Topologies
 - > Multi-bank L2$ Provides Ample Snoop Bandwidth

Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

Multi-chip Interconnect

- Distributed Points of Global Ordering
 - > Coherence Planes Operate Independent of One Another
 - > Physical Address Conflicts Serialized by Coherence Hubs
- Cache States
 - > MOESI States Tracked by L2$
 - > C2C Transfers on Snoop Hits to M, O, E and S (MCU Node) States
- Independent Packet Processing Between 3 Virtual Channels
- OOO Snoop Responses
 - > Fairness Algorithm
 - > Progress Beyond Stalled Snoops



Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

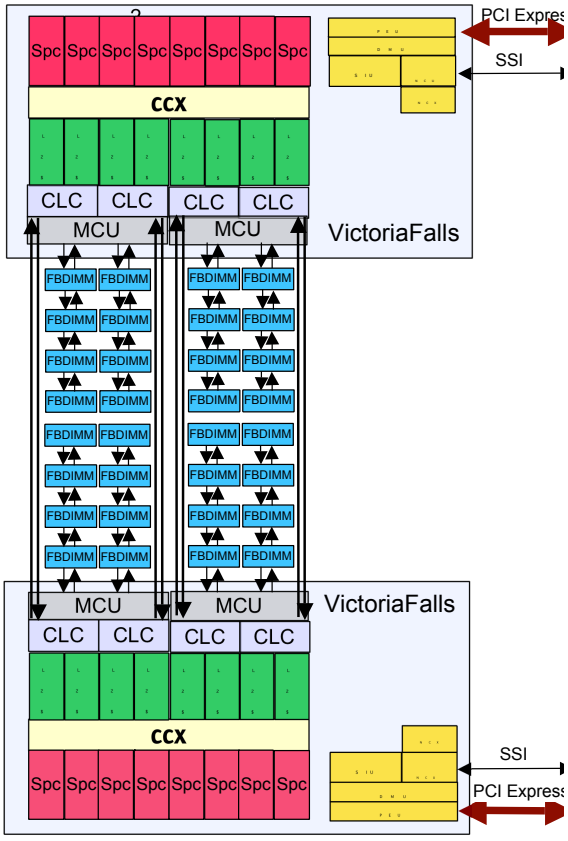
Multi-chip Interconnect

- Physical Layer
 - 14 Transmit and 14 Receive Lanes per Channel
 - 4.8GT/sec, 8.4 GB/sec Raw Link Bandwidth per Direction
 - Clock Recovery, Bit/symbol Alignment, Initialization and Training
- Reliable Point-to-point Data Link Layer
 - Fixed Size Frames: 144bit Payload + 24bit CRC
 - 7.2GB/sec Payload BW, 6.4GB/sec Peak Data BW per Link, Full-Dupl
 - Hardware Frame Replay, Link Retrain and Lane Failover with Full CR Continuance
- Virtual Channels
 - Request/Request-ack
 - Response
 - Data/Critical Data
- Transaction Layer
 - 3/7/18-Byte Packets, May Cross Frame Boundaries
 - Weighted Round Robin Arbitration of Egress Virtual Channel Queues

Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

Dual-chip Architecture

- 16 SPARC Cores
 - 128 Threads
- Coherent Interconnect
 - 4 Coherence Links, Full-Duplex
 - 65GB/sec Raw Bisection
- 8 FBDIMM Channels
 - 42GB/sec (Theoretical Peak) Read
 - 21GB/sec (Theoretical Peak) Write
 - DDR2-667
- Integrated I/O Bridges
 - 2 x8 Lane PCI Express Ports @ 2.5GT/sec per Lane Full Duplex
 - 1024 Concurrent IO Address Translations (Virtual-to-Real, Real-to-Physical)
 - Relaxed DMA Ordering



Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

Dual-chip Local Memory Coherent Read

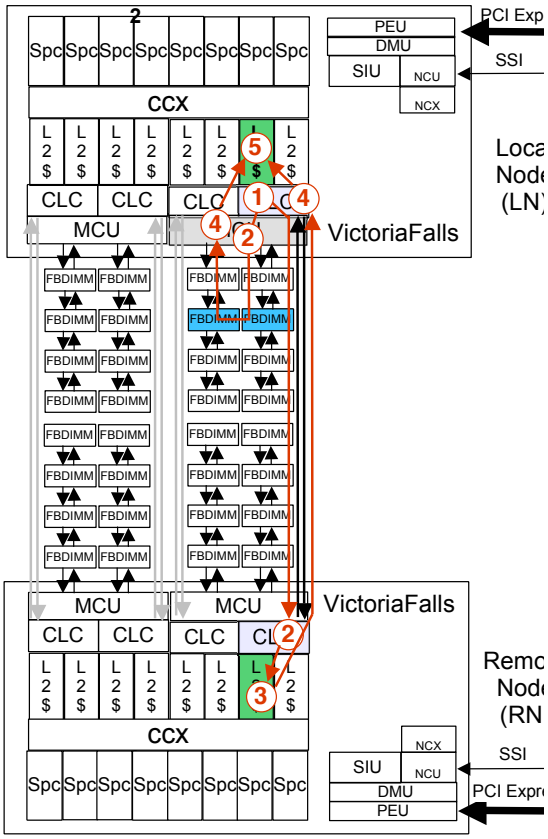
1 LN Coherence Hub serialization. Forwards read request to LN MCU & snoop request to RN CLC.

2 LN MCU FBDIMM access. RN CLC forwards snoop request to RN L2$ bank.

3 RN L2$ bank snoop operation. Returns snoop response and potentially C2C data to LN CLC.

4 LN CLC forwards snoop response, memory and C2C data to LN L2$ bank as they arrive.

5 LN L2$ bank bypasses resolved return data to upstream cache and fills allocated L2$ entry.



Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

Dual-chip Remote Memory Coherent Read

1 LN CLC forwards read request to RN CLC.

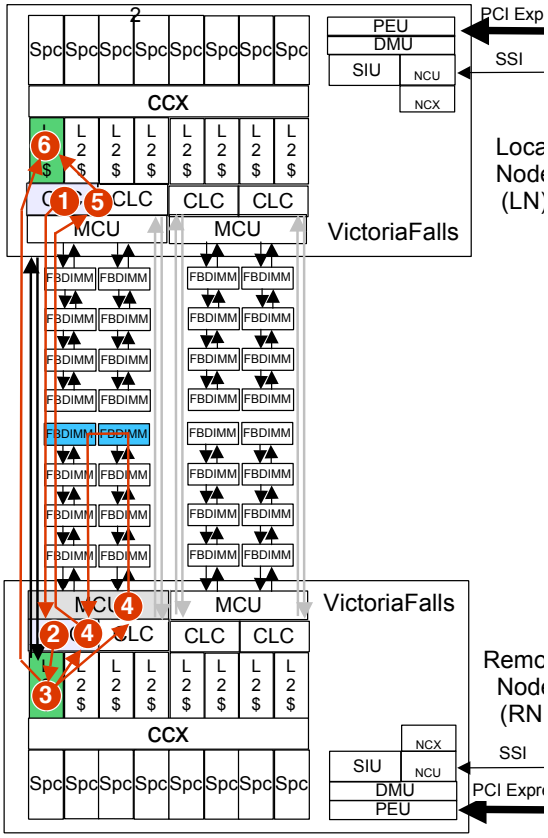
2 RN Coherence Hub serialization. RN CLC forwards Request-ack to LN L2$ Bank and snoop request to RN L2$ Bank.

3 RN L2$ bank snoop operation, returns snoop status result to RN CLC to activate read request to RN MCU.

4 LN MCU FBDIMM access if no L2$ copyback. RN CLC forwards snoop response and memory or copyback data to LN CLC.

5 LN CLC forwards snoop response and return data to local L2$ bank as they arrive.

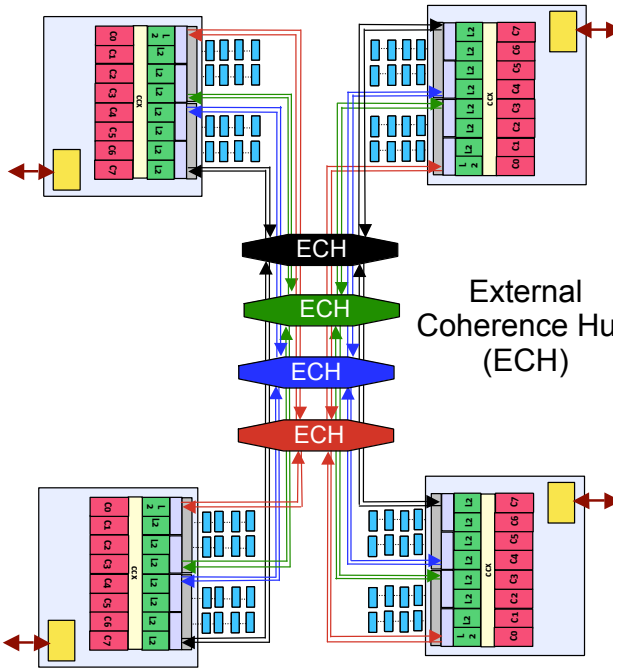
6 L2$ forwards/bypasses resolved return data to upstream cache and fills L2$.



Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

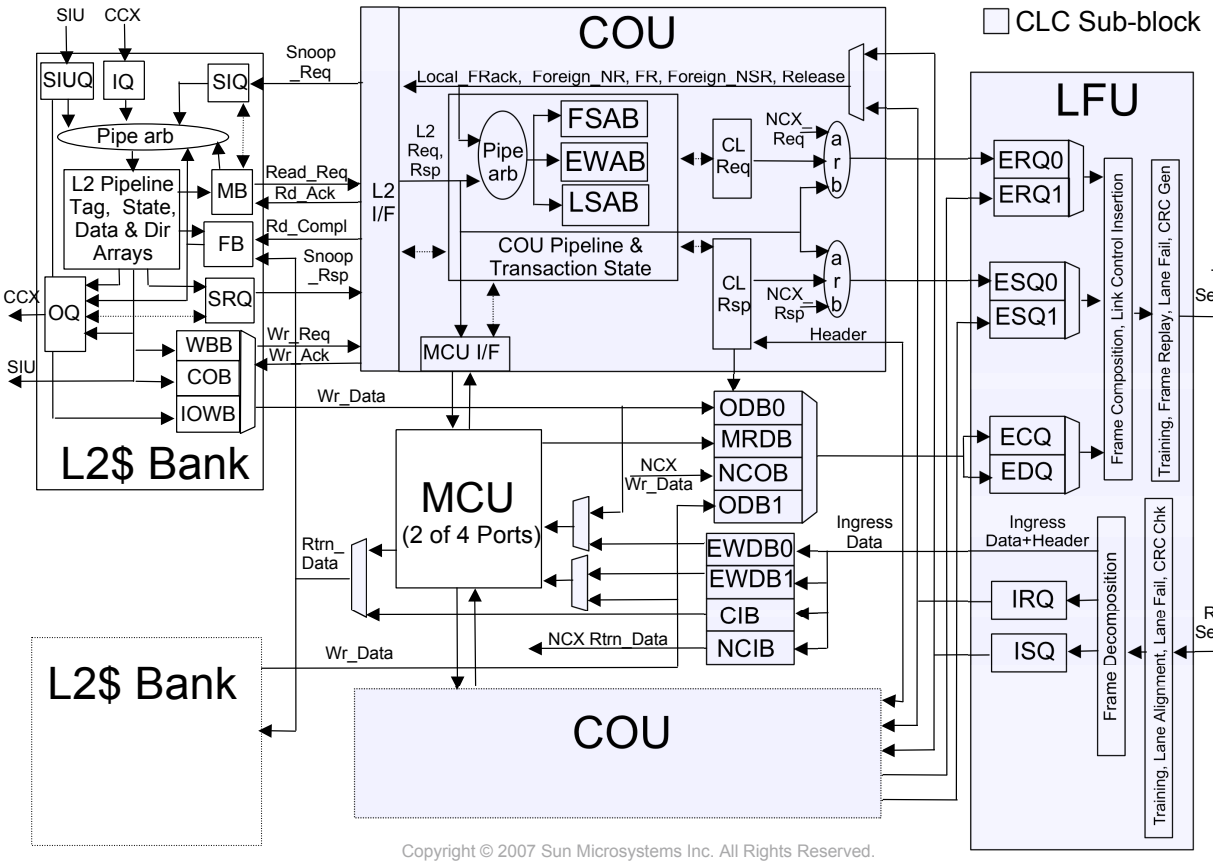
Quad-chip Architecture

- 32 SPARC Cores
 - > 256 Threads
- Coherent Interconnect
 - > 4x4 Coherence Links, Full-Duplex
 - > 130GB/sec Raw Bisection
- 16 FBDIMM Channels
 - > 84GB/sec (Theoretical Peak) Read
 - > 42GB/sec (Theoretical Peak) Write
- 4 Integrated I/O Bridges
- 4 External Coherence Hubs
 - > Global Ordering
 - > Snoop Response Aggregation
 - > Data Filtering
 - > Destination Flow Control



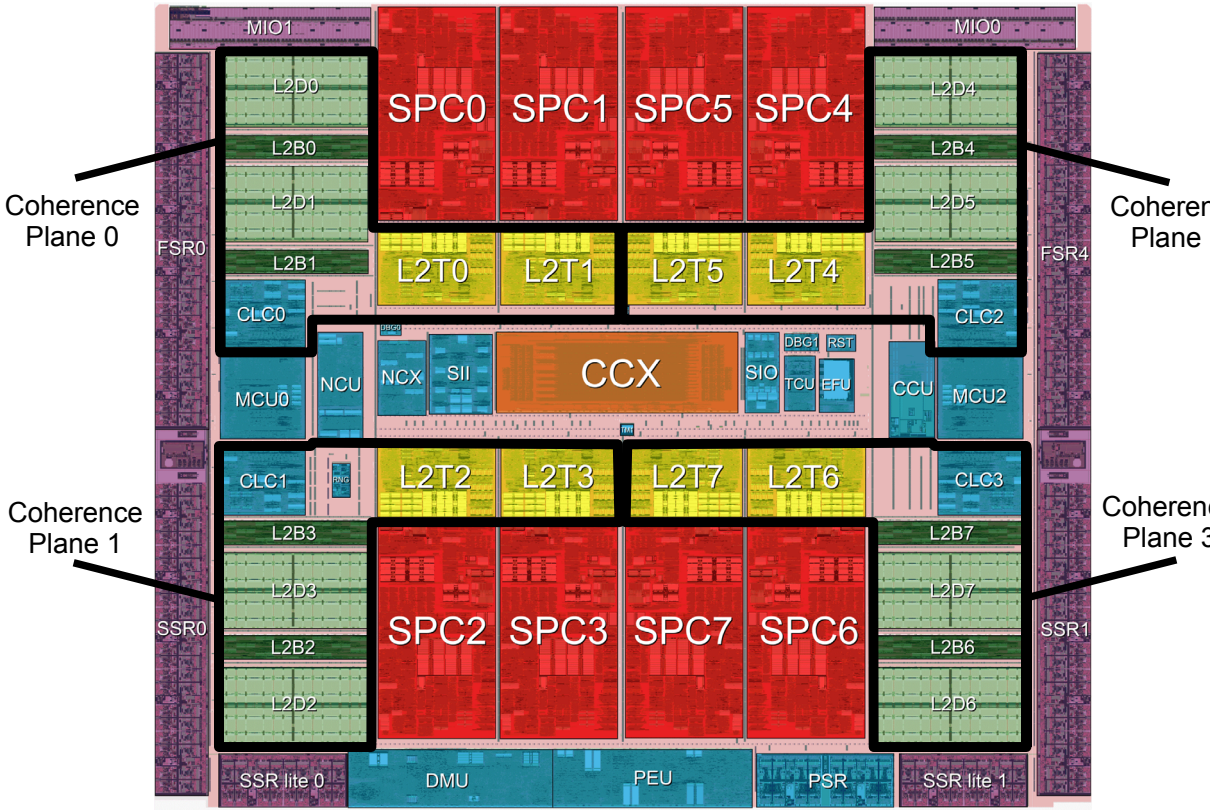
Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

Coherence Plane Microarchitecture



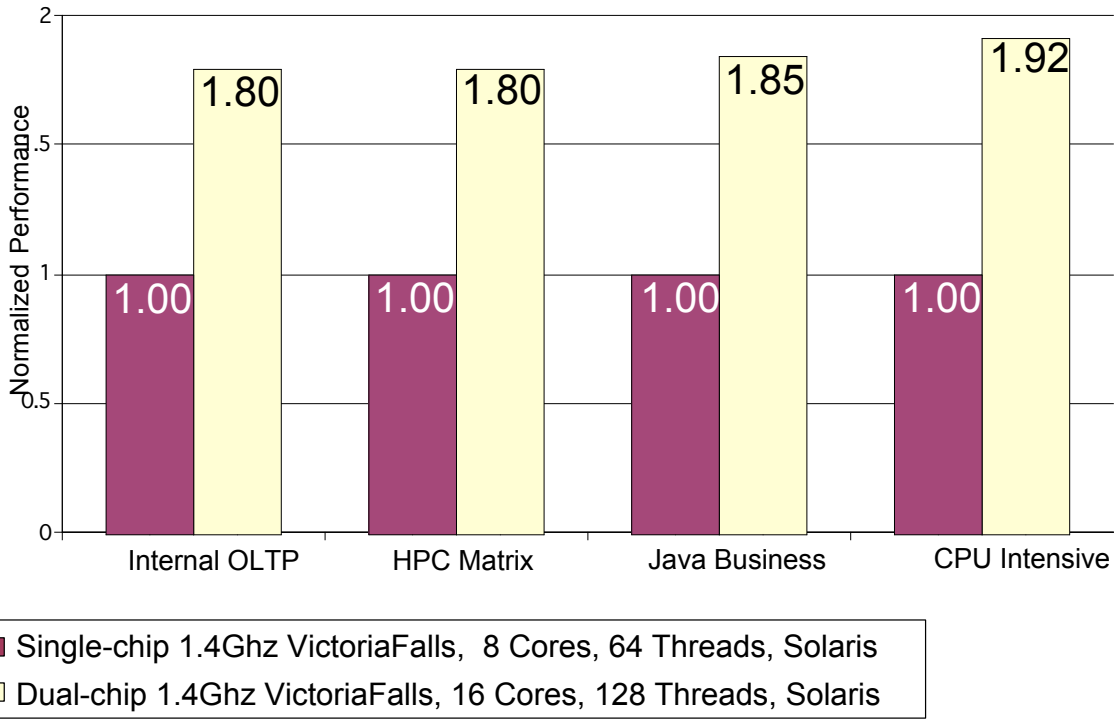
Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

Coherence Plane Layout



Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

Performance Scaling



Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

Summary

- VictoriaFalls Scales Throughput Performance via Multi-bank Caches, Multi-port Memory Controllers and Multi-plane Coherent Interconnects.
- VictoriaFalls is Comparable to the Niagara2 Chip Multiprocessor in Terms of Die Area and Power Envelope.
- VictoriaFalls Enables Large (128 to 256) Thread Count, High Capacity and High Throughput Performance in Dense, Power-Efficient Form Factors.

Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

Acknowledgements

- Ramaswamy Sivaramakrishnan
- Jeffrey Oplinger
- Sumti Jairath
- Sebastan Turullols
- Damien Walker
- William Bryg
- Rick Hetherington
- Connie Cheung
- Robert Dickson
- Rodrigo Liang
- Vida Ghodssi
- Kunle Olukotun
- David Greenhill
- Curtis McAllister
- David Smentek
- Ali Vahidsafa
- Tom Karabinas
- Denis Sheahan
- Greg Grohoski
- Robert Golla
- Lawrence Spracklen
- James Laudon

Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.

Thank you ...

stephen.phillips@sun.com

Microarchitecture Terminology

- L2$ Bank – 8 Instances
 - IQ: Core Input Queue
 - SIUQ: System (DMA) Input Queue
 - SIQ: Snoop Input Queue
 - SRQ: Snoop Response Queue
 - OQ: Output Queue
 - MB: Miss Buffer
 - FB: Fill Buffer
 - WBB: Writeback Buffer
 - COB: Copyback Buffer
 - IOWB: I/O Writeback Buffer
- Link Framing Unit (LFU)
 - ERQ0, ERQ1: Egress Request Queues
 - ESQ0, ESQ1: Egress Status Queues
 - ECQ: Egress Critical Data Queue
 - EDQ: Egress Data Queue
 - IRQ: Ingress Request Queue
 - ISQ: Ingress Status Queue
- Coherence & Ordering Unit (COU) – 8 Instances
 - LSAB: Local Snoop Address Buffer
 - FSAB: Foreign Snoop Address Buffer
 - EWAB: External Writeback Address Buffer
- CLC Datapath – 4 Instances
 - ODB0, ODB1: Output Data Buffers
 - MRDB: Memory Return Data Buffers
 - NCOB: Non-Cacheable Output Buffer
 - EWDB0, EWDB1: External Writeba Data Buffers
 - CIB: Copyback Input Buffer
 - NCIB: Non-Cacheable Input Buffer

Copyright © 2007 Sun Microsystems Inc. All Rights Reserved.